Goal: Smoother Density Estimation.

3 Kernel Density Estimation

Recall the definition of a density function

$$f(y) \equiv \frac{d}{dy}F(y) \equiv \lim_{h\to 0} \underbrace{\frac{F(y+h)-F(y-h)}{2h}} = \lim_{h\to 0} \underbrace{\frac{F(y+h)-F(y)}{h}}_{\text{take this and approximate w/ small fixed } h\text{, use Ecdf.}},$$

where $F(x)$ is the cdf of the random variable Y.

$$\hat{f}(x) = \frac{\hat{F}_n(y+h) - \hat{F}_n(y)}{h} \quad \text{histogram.}$$

What if instead, we replace $\underbrace{F(y+h) - F(y-h)}_{\text{with ecdf}}$?

$$\Rightarrow \hat{f}(y) = \frac{\hat{F}_n(y+h) - \hat{F}_n(y-h)}{2h} = \frac{\#\{y_i \in (y-h, y+h]\}}{2nh}$$

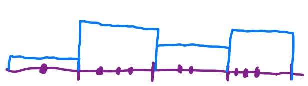
$$= \frac{\sum_{i=1}^{n} \mathbb{I}(y_i \in (y-h, y+h]\}}{2nh}$$

$$= \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2}\mathbb{I}(y-h < y_i \le y+h)$$

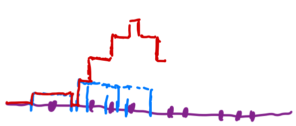
$$= \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2}\mathbb{I}\left(-1 < \frac{y-y_i}{h} \le 1\right)$$

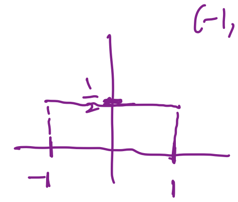
$$= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right) \text{ where } K \text{ is a Uniform density on } (-1,1]$$

Kernel function.



histogram





still not continuous (because Uniform density not continuous!)

$\Rightarrow$ another Kernel may lead to smoother estimate.

A kernel function assigns weights to the contribution given by each y_i to $\hat{f}(y)$, depending on proximity to y.

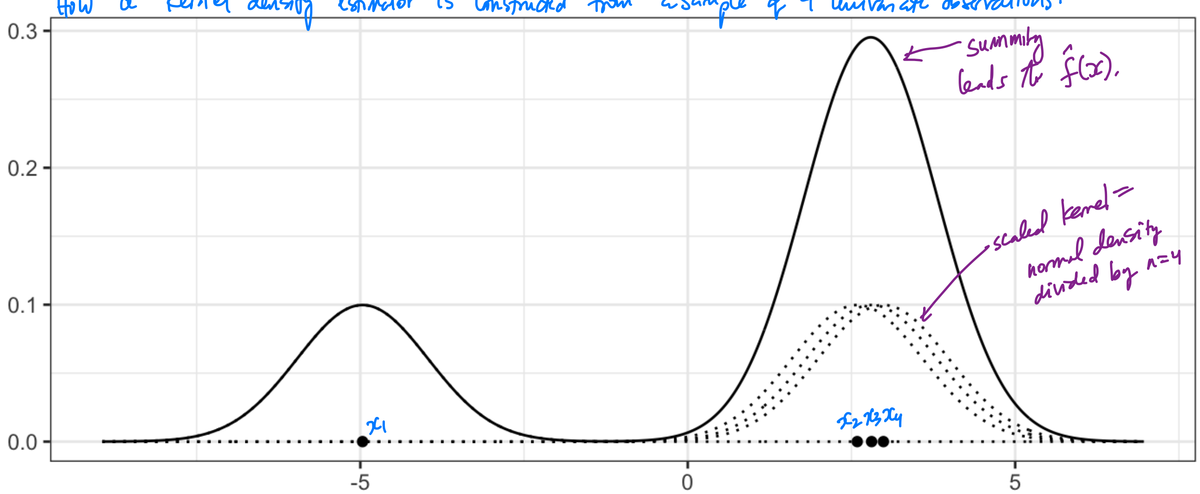
This will weight all points within h of y equally. A univariate *kernel density estimator* will allow a more flexible weighting scheme.

Typically, kernel functions are positive everywhere and symmetric about zero.

Examples of ideas for such functions? Normal density, Student t (others exist).

Additionally, constraining K so that $\int z^2 K(z)\,dz = 1$ allows h to play role of scale parameter (not required).

How a kernel density estimator is constructed from a sample of 4 univariate observations:



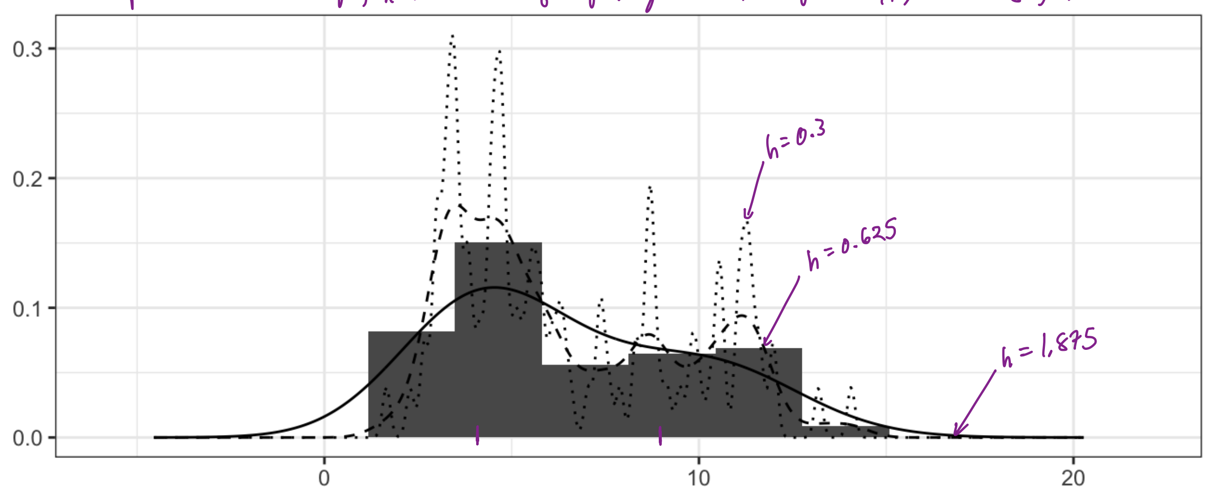
3.1 Choice of Bandwidth

The bandwidth parameter controls the smoothness of the density estimate. for a given kernel.

bandwidth determines tradeoff btw/ bias and variance.

The tradeoff that results from choosing the bandwidth + kernel can be quantified through a measure of accuracy of $\hat{f}$, such as MISE.

Example: Bimodal density, n=100 from equally weighted mixture of N(4,1) and N(9,4).



For large h, oversmoothing (lose 2nd mode).

For small h, undersmoothing (many false modes).

(AMISE)

To understand bandwidth selection, let us analyze MISE. Suppose that K is a symmetric, continuous probability density function with mean 0 and variance $0 < \sigma_K^2 < \infty$. Let $R(g) = \int g^2(z)dz$. Recall that

$$\text{MISE}(h) = \int \text{MSE}(\hat{f}(x))dx = \int var\{\hat{f}_h(x)\} + [bias\{\hat{f}_h(x)\}]^2 dx$$

Now let $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

Bias: Note $E\{\hat{f}(x)\} = \frac{1}{h}\int K\left(\frac{x-u}{h}\right) f(u)du$

$$= \int K(t) \underbrace{f(x-ht)} dt \quad \text{(change of variable).}$$

and using Taylor's expansion, $f(x-ht) = f(x) - htf'(x) + \frac{h^2t^2}{2}f''(x) + o(h^2)$

$$\Rightarrow E\{\hat{f}(x)\} = f(x) + \frac{h^2\sigma_K^2}{2}f''(x) + o(h^2) \quad \leftarrow \text{because } K \text{ is symmetric about } 0.$$

$$\text{so, } [bias\{\hat{f}(x)\}]^2 = \frac{h^4\sigma_K^4}{4}[f''(x)]^2 + o(h^4).$$

$$\Rightarrow ISB = \int [bias\{\hat{f}(x)\}]^2 dx = \frac{h^4\sigma_K^4}{4}R(f'') + o(h^4).$$

Variance: $Var\{\hat{f}(x)\} = \frac{1}{n}Var\left\{\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right\}$

change of variable $= \frac{1}{nh}\int K(t)^2 f(x-ht)dt - \frac{1}{n}\left[E\left\{\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right\}\right]^2$

Taylor's expansion $= \frac{1}{nh}\int K(t)^2[f(x) + o(1)]dt - \frac{1}{n}[f(x) + o(1)]^2$

$$= \frac{1}{nh}f(x)R(K) + o\left(\frac{1}{nh}\right)$$

$$\Rightarrow IV = \int var\{\hat{f}(x)\}dx = \frac{R(K)}{nh} + o\left(\frac{1}{nh}\right).$$

$$\text{and } MISE = \underbrace{\frac{R(K)}{nh} + \frac{h^4\sigma_K^4R(f'')}{4}}_{AMISE} + o\left(\frac{1}{nh} + h^4\right).$$

To minimize AMISE with respect to h, seek value of h that avoid excessive bias and variance.

optimal bandwidth $h_0 = \left(\dfrac{R(K)}{n \sigma_K^4 R(f'')} \right)^{1/5}$

$\Rightarrow$ minimal AMISE: $AMISE_0 = \frac{5}{4} \left[\sigma_K R(K) \right]^{4/5} R(f'')^{1/5} n^{-4/5}$

Recall for histogram, $AMISE_0 = \left[\dfrac{R(f')}{16} \right]^{1/3} n^{-2/3}$

$\Rightarrow$ w kernel estimate, getting closer to parametric rate of n^{-1}

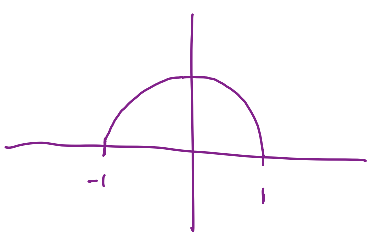
The term $R(f'')$ measures the roughness of the true underlying density. In general, rougher densities are more difficult to estimate and require smaller bandwidth.

The term $[\sigma_K R(K)]^{4/5}$ is a function of the kernel function K.

We could choose K to minimize $\left[\sigma_K R(K) \right]^{4/5}$:

If K restricted to be a proper density (w/ some moment conditions), minimizer is scaled version of a quadratic density.

$K(u) = \frac{3}{4} (1 - u)^2 \, \mathbb{I}(|u| \leq 1)$



"Epanechnikov kernel" (more later).

3.1.1 Cross Validation - Want to evaluate quality (Q) of $\hat{f}$ without using data twice (once for fitting $\hat{f}$, once for evaluating).

$\Rightarrow$ again use $\hat{f}_{-i}(X_i) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)$ and let $\hat{Q}(h)$ be a function of $\hat{f}_{-i}(X_i)$ that assesses quality.

e.g. ISE from before.

If we choose $\hat{Q} = ISE$, choose h to minimize $R(\hat{f}) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)$

could instead choose $\hat{Q}(h)$ as the pseudo likelihood $PL(h) = \prod_{i=1}^{n} \hat{f}_{-i}(X_i)$ and choose h to maximize!

Typically undersmoothed $\Rightarrow$ too bumpy.

3.1.2 Plug-in Methods

If the reference density f is Gaussian and a Gaussian kernel K is used,

$$h_0 = 1.059\, \sigma\, n^{-1/5}$$

sample variance, or IQR plugin approach.

Typically oversmoothed

Empirical estimation of $R(f'')$ may be a better option. Recall $h_0 = \left(\frac{R(K)}{n \sigma_K^4 R(f'')}\right)^{1/5}$

We could use a kernel density estimator for f'':

$$\hat{f}''(x) = \frac{d^2}{dx^2} \left\{ \frac{1}{n h_1} \sum_{i=1}^{n} L\left(\frac{x - x_i}{h_1}\right) \right\}$$

$$= \frac{1}{n h_1^3} \sum_{i=1}^{n} L''\left(\frac{x - x_i}{h_1}\right)$$

where L is a sufficiently differentiable kernel and h_1 = bandwidth to estimate f''.

Note: estimating f and f'' (for $R(f'')$) will require different bandwidths.

2 stage approach [Sheather-Jones method]:

1. Use Gaussian plug in for h_1. This bandwidth is used to estimate $R(f'')$.
2. h is calculated using $h = \left(\frac{R(K)}{n \sigma_K^4 \hat{R}(f'')}\right)^{1/5}$ used to produce the final kernel density estimate.

3.2 Choice of Kernel

There are two choices we have to make to perform density estimation:

Kernel and bandwidth.

The shape of Kernel is much less important than bandwidth.

3.2.1 Epanechnikov Kernel

Recall $AMISE_0 = \frac{5}{4}\left[\sigma_K R(K)\right]^{4/5} R(f'')^{1/5} n^{-4/5}$.

The *Epanechnikov kernel* results from choosing K to minimize $[\sigma_K R(K)]^{4/5}$, restricted to be a symmetric density with finite moments and variance equal to 1

$\Rightarrow K(u) = \frac{3}{4}(1-u^2)\mathbb{I}(|u|\le 1)$

and. $\sigma_K R(K) = \frac{3}{5\sqrt{5}}$.

$\Rightarrow$ The ratio of $\frac{\sigma_K R(K)}{3/5\sqrt{5}}$ provides a measure of relative inefficiency of other Kernels.

$\downarrow$

multiplicative factor for equivalent sample Size needed to achieve same AMISE.

Kernel	Inefficiency	
Epanechnikov	1	
Gaussian	1.0513	
Uniform	1.0758	
Biweight	1.0061	$\frac{15}{16}(1-u^2)^2$
Triweight	1.0135	$\frac{35}{32}(1-u^2)^3$

kernel choice doesn't make much difference!

3.2.2 Canonical Kernels

Unfortunately a particular value of h corresponds to a different amount of smoothing depending on which kernel is being used.

Let h_K and h_L denote the bandwidths that minimize AMISE when using symmetric kernel densities K and L. Then,

$$\frac{h_K}{h_L} = \frac{\left(R(K)/\sigma_K^4\right)^{1/5}}{\left(R(L)/\sigma_L^4\right)^{1/5}} = \frac{\delta(K)}{\delta(L)}.$$

$\Rightarrow$ to change from bandwidth h for kernel K to a bandwidth that gives an equivalent amount of smoothing for L, use bandwidth $h\,\delta(L)/\delta(K)$.

Epanechnikov: $\delta(K)\ 15^{1/5}$, Gaussian: $\delta(K) = \left(\frac{1}{\sqrt{2\pi}}\right)^{1/5}$, Uniform: $\delta(K) = \left(\frac{9}{2}\right)^{1/5}$.

Suppose we rescale a kernel shape so that $h = 1$ corresponds to a bandwidth of $\delta(K)$,

The kernel density estimator can be written as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h\delta(K)}(x - x_i) \quad \text{where} \quad K_{h\delta(K)}(z) = \frac{1}{h\delta(K)} K\left(\frac{z}{h\delta(K)}\right).$$

$K_{\delta(K)}$ is called a "canonical kernel"

$\downarrow$

benefit: a single value of h can be used interchangeably for each canonical kernel without affecting the amount of smoothing.

For a canonical kernel w/ bandwidth h,

$$\text{AMISE} = \left(\sigma_K R(K)\right)^{4/5} \underbrace{\left(\frac{1}{nh} + \frac{h^4 R(f'')}{4}\right)}$$

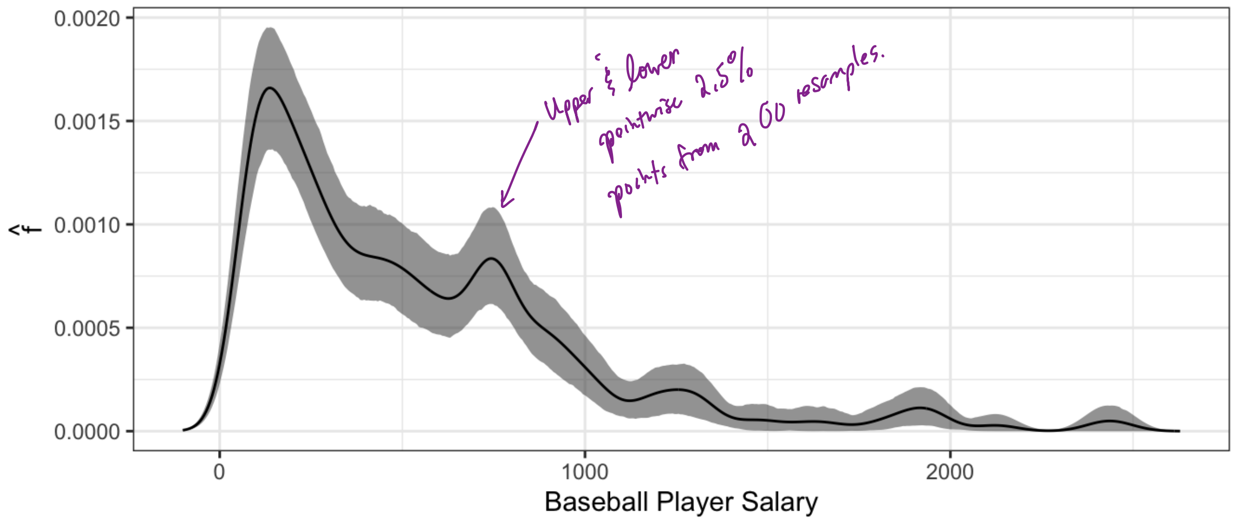
bias/variance trade-off no longer confounded w/ $R(K)$ (choice of kernel).
AND optimal kernel choice doesn't depend on bandwidth.

So, the Epanechnikov kernel shape is optimal for any desired degree of smoothing.

3.3 Bootstrapping and Variability Plot

$\underbrace{\text{Variability}}_{\text{describing uncertainty in } \hat{f}}$

1. A sample of size n is drawn w/ replacement from data.
2. bandwidth chosen for new sample based on Sheather-Jones method, density estimate determined.
3. Repeat 1-2 many times w/ values of $\hat{f}$ recorded at a fixed grid of values.



Note: This is NOT a 95% CI for f, but a representation of the variability in the process of estimating $\hat{f}$.

↳ can't say anything about $P(\hat{f}(x)_{lower} \le f(x) \le \hat{f}(x)_{upper})$ because bias.

widens at peaks and valleys, narrows where $\hat{f}$ more flat.

↳ consistent w/ MSE of $\hat{f}(x)$ directly related to $[f''(x)]^2$